

# How to infer relative fitness from a sample of genomic sequences

Adel Dayarian<sup>1</sup> and Boris I. Shraiman<sup>1,2</sup>

<sup>1</sup>*Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA*

<sup>2</sup>*Department of Physics, University of California, Santa Barbara, CA*

Mounting evidence suggests that natural populations can harbor extensive fitness diversity with numerous genomic loci under selection. It is also known that genealogical trees for populations under selection are quantifiably different from those expected under neutral evolution and described statistically by Kingman's coalescent. While differences in the statistical structure of genealogies have long been used as a test for the presence of selection, the full extent of the information that they contain has not been exploited. Here we shall demonstrate that the shape of the reconstructed genealogical tree for a moderately large number of random genomic samples taken from a fitness diverse, but otherwise unstructured asexual population can be used to predict the relative fitness of individuals within the sample. To achieve this we define a heuristic algorithm, which we test *in silico* using simulations of a Wright-Fisher model for a realistic range of mutation rates and selection strength. Our inferred fitness ranking is based on a linear discriminator which identifies rapidly coalescing lineages in the reconstructed tree. Inferred fitness ranking correlates strongly with the actual fitness, with top 10% ranked being in the top 20% fittest with false discovery rate of 0.1-0.3 depending on the mutation/selection parameters. The ranking also enables to predict the common genotype of the future population. While the inference accuracy increases monotonically with sample size, sample sizes of 200 nearly saturate the performance. We propose that our approach can be used for inferring relative fitness of genomes obtained in single-cell sequencing of tumors and in monitoring viral outbreaks.

## I. INTRODUCTION

Most of mutations have minimal effects on the fitness of the organism and much of the analysis of the genomic data on populations (see [1] for a review of methods) has been based on the neutral hypothesis, according to which the dynamics of genetic polymorphisms and the overall genetic diversity of the population are governed by the neutral *drift*, i.e. stochastic fluctuations in the number of offspring. According to the neutral picture, deleterious mutations are eliminated by selection fast enough to not significantly contribute to population diversity and beneficial mutations are rare enough to produce only occasional adaptive *sweeps*, where the population is taken over by the offspring of the adaptive genotype, transiently suppressing neutral genetic diversity. Statistical properties of genealogies generated by neutral dynamics in asexual populations are understood in great detail [2] in terms of the Kingman's *coalescent* process [3] which follows the ancestors of the present population back in time as far as the *Most Recent Common Ancestor* (MRCA). The neutral coalescent [2] forms the basis for estimating mutation and recombination rates and provides the null hypothesis in tests for the presence of selection [4, 5].

Yet, as advances in sequencing have made it possible to obtain quantitative data on genetic diversity, numerous studies have reached the conclusion that non-neutral polymorphisms are ubiquitous in populations across the spectrum of life: from viruses [6–8] and bacteria [9] to flies [10], from mitochondria [11] to cells in cancerous tumors [12]. In addition, laboratory evolution experiments in bacteria [13, 14] and yeast [15, 16] have demonstrated directly that large asexual populations contain numerous sub-clones that are continuously generated by mutation

and compete for fixation. Thus, large asexual population cannot be assumed selectively neutral.

Since the MRCA in a fitness diverse population was with high probability among the very fittest of its generation, the dynamics of genealogical coalescence is controlled by the time it takes for surviving lineages to converge, as they are tracked back in time, on the leading edge of the fitness distribution. The time scale for this genetic coalescence is set by the fitness differential between the best and the typical genotypes - it is in essence the *genetic turnover* time, or the time it takes for the offspring of the best genotype to take over the population. This delays coalescence, giving genealogical trees for populations under selection a "comb-like" appearance that is strikingly different from the neutral case [11, 17].

An excellent illustration of the "genealogical anomalies" - i.e. large deviations from neutral genealogical structure [17] - is provided by the very interesting recent study [11] of mitochondrial diversity in three distinct populations of whale lice, *Cyamus ovalis*, where the authors demonstrate that the observed genealogies are statistically consistent with a non-neutral model with frequent mutations of small selective effect.

Our analysis will be based on a similar model of asexual evolutionary dynamics driven by small deleterious and beneficial mutations. In Fig. 1 we show schematically a sample of continuous genealogy of a fixed size population governed by the Wright-Fisher dynamics [2] incorporating genetic drift, mutation and natural selection. The example in Fig. 1 covers the period over which the offspring of one of the genomes in the top population take over the whole population at the bottom of the figure. We ask, given a sample of genomes from the "present time" population (shown in Fig. 1 as red discs), can one predict

genetic future of the population? Or more specifically, can one identify within the present sample the closest relatives of the future population: i.e. individuals that are on, or closest to, the genealogical backbone of the future population? Since long term survival is correlated with fitness, this task is closely related to the problem of identifying the fitter fraction of the present day sample.

Below, we shall demonstrate that the anomalous structure of the genealogical tree reconstructed for the sample of genomes can serve not only as the evidence for the action of selection, but also as the basis of inference of the relative fitness of sampled individuals and their sequence closeness to the fittest genomes. Information pertinent to this inference is contained in the pattern of coalescence experienced by different lineages. In the nutshell: lineages which undergo a lot of coalescence much before others, are relatively fit, while the lineages which do not merge until after the arrival of the coalescence “peloton” are less fit. Our study builds on the considerable recent progress in the theoretical understanding of natural selection and drift dynamics in fitness-diverse asexual populations [18–25] and the emerging description of corresponding genealogies in terms of the *Bolthausen-Sznitman Coalescent* (BSC) [11, 22, 23, 26–30]. We shall focus on the asexual case and address how the approach might be extended to the analysis of recombining populations in the discussion.

After formulating the model, we shall i) provide examples of genealogies illustrating their anomalous shape as compared to the neutral coalescent and ii) demonstrate the correlation between ancestral *weight* - the fraction of the present day sample constituted by its offspring - and the mean fitness of the latter. We shall then define a fitness-ranking “score” based on the suitably integrated ancestral weights along the reconstructed lineage of each individual in the sample. Applying the ranking to numerous sampling realizations (for populations with the same and with different mutation/selection parameters) and comparing each realization to the true fitness known from the forward simulation, we demonstrate the ability of the proposed algorithm to infer the relative fitness of sampled genomes and to identify genotypes that are likely to survive into the future. The Discussion will address possible applications and generalizations of the proposed inference method.

## II. MODEL

Consider an asexual population of size  $N$  that evolves with non-overlapping generations under the influx of deleterious and beneficial mutations. New mutations arise at the rate  $\mu + \mu_0$  (per genome per generation) with a fraction  $\epsilon\mu$  being beneficial,  $(1 - \epsilon)\mu$  deleterious and the remainder  $\mu_0$  being neutral. For simplicity we assume both beneficial and deleterious mutations to have the same effect size  $s \ll 1$  changing the fitness additively:  $F_i \rightarrow F_i \pm s$ . As in the Wright-Fisher model, natural se-

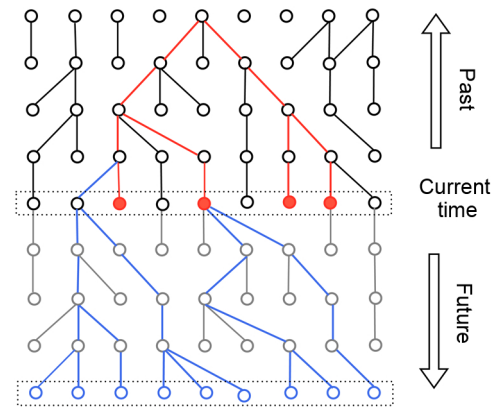


FIG. 1. Schematic example of a genealogical trajectory, from past into the future, of an asexual population with fixed size ( $N = 9$ ) and non-overlapping generations. Nodes represent individual genomes, each linked to its ancestor in the previous generation. The example illustrates coalescence of the lineages of the bottom population towards its MRCA within the top population. The genealogical tree of a random sample (red) from the “current time” population partially overlaps the genealogy of the future population (blue). While actual ancestors of the future population (shown in blue) may or may not fall into the current sample, one can still define sample members that are closest to the surviving lineages. Identifying close relatives of future populations is the goal of our study.

lection acts by biasing the probability of an individual genome to appear in the next generation, which is taken to be proportional to  $\exp(f_i)$  with  $f_i = F_i - \bar{F}$  being the individual fitness relative to the mean fitness of the population  $\bar{F}$ , which in general is a function of time.

We carried out simulations for  $N = 64,000$  and several plausible parameter combinations in the range of  $\mu = 10^{-4} - 10^{-2}$  and  $s = 10^{-3} - 10^{-2}$ , with  $\epsilon = 0.1$  and  $0$  and  $\mu_0 = 10\mu$ . The genealogical trees were constructed in two ways. We recorded the genealogies in the course of the forward simulation, providing exact ancestries of any sample in the population. In addition, an inferred genealogy of random samples (between 30–500 genomes) was constructed using standard neighbor joining/UPGMA-derived methods detailed in the Supplementary Information (SI).

## III. RESULTS

In the parameter range considered, simulated population exhibit substantial fitness diversity  $\sigma \approx 10^{-3} - 10^{-2}$  arising from about  $10 - 10^3$  simultaneously segregating non-neutral polymorphisms. Fig. 2A-B shows examples of the population-wide fitness distribution for two different mutation rates (see SI for additional examples). In general, the genetic diversity in the population is an increasing function of  $\mu/s$ . For the highest mutation rate and lowest selection coefficients considered,  $\mu = 10^{-2}$

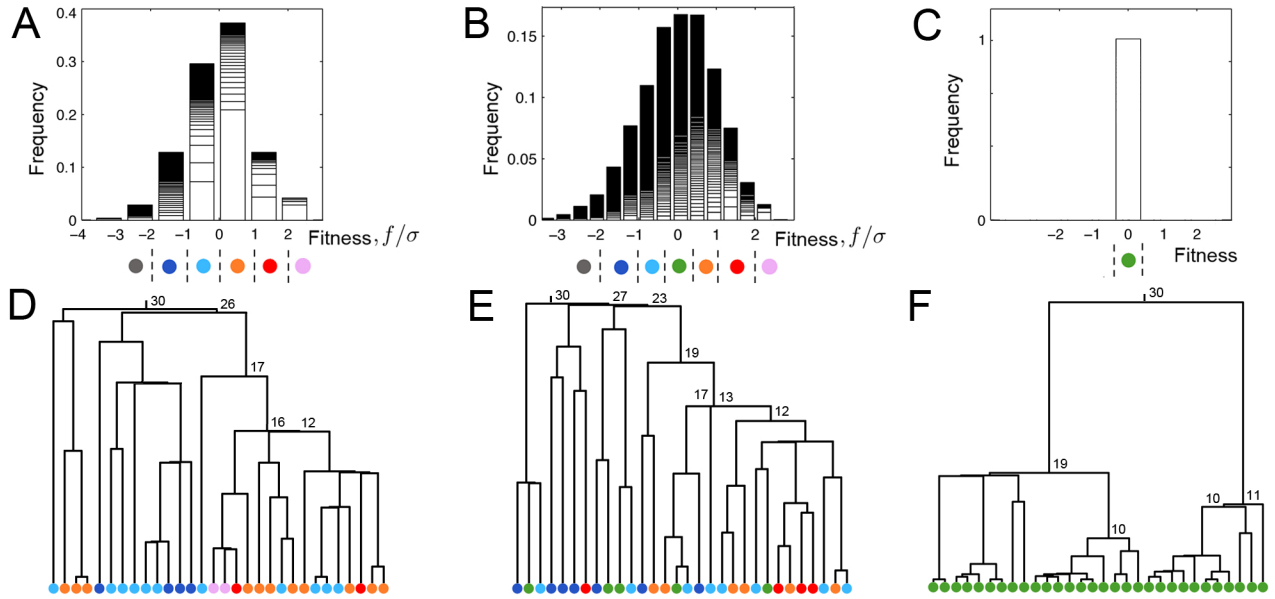


FIG. 2. Fitness distributions and examples of genealogical trees. (A) Fitness distribution at one time point for a population with  $\mu = 10^{-3}$ ,  $s = 2 * 10^{-3}$ . Each bin corresponds to a fitness class and each class is composed of multiple clones. (Here, clones are defined using only the non-neutral mutations.) Also shown is the color-code used in (D). (B) Same as (A) but for a higher mutation rate  $\mu = 10^{-2}$ . (C) Same as (A) but for a neutral population. (D) A typical genealogical tree for a random sample of size  $n = 30$  from the same population as (A). Each circle corresponds to one sampled genome and the color represents its fitness. Branch lengths are drawn in linear proportion to the corresponding time interval. Numbers next to internal nodes are the weights of the corresponding ancestors (only weights  $\geq 10$  are shown). Note the striking asymmetry of branching, as the weight decreases in small steps along the lineage with weights marked. (E) Same as (D) but for the population shown in (B). Note that the colors (grey and plum) corresponding to the extremes of the distribution (B) are absent from the small sample shown. (F) Same as (D) but for a neutral population. Note the short terminal legs and more symmetric branching.  $N = 64000$  and  $\epsilon = 0.1$  for all the panels.

and  $s = 10^{-3}$ , the population is in a weak selection regime corresponding to high genetic diversity, Fig. 2B. Lower mutation rates, as in Fig. 2A, exhibit a more clonal structure: evolutionary dynamics in this regime can be thought of as competition between multiple mutant clones.

Fig. 2D and E show typical examples of genealogical trees constructed for random samples of size  $n = 30$  drawn from the populations corresponding to Fig. 2A and B, respectively. The fitness of sampled genomes, which we know from the forward simulation, is visualized using color. Also shown are ancestral weights along some of the lineages. This weight,  $w_i$ , is defined as the number of genomes in the sample which are direct descendants of genome  $i$ . For example, each leaf at the bottom carries weight  $w=1$ , while the node at the top carries the full weight of the sample  $n = 30$ . For the sake of comparison, we have also shown a typical genealogical tree for a neutrally evolving population in Fig. 2F.

#### A. Distortion in the shape of trees in the presence of selection

One immediately notes two striking (and well known [11, 17]) differences distinguishing Fig. 2D-E and F: Fit-

ness diverse populations i) have long terminal legs and are compressed towards the MRCA root of the tree, ii) exhibit strong asymmetry of branching. These anomalies are quantified in Fig. 3. Fig. 3A presents distributions of pairwise coalescent times in the population,  $\tau_{ij}$ , for  $\{i, j\}$  genome pairs for several parameter sets. In the Kingman's coalescent,  $\tau_{ij}$  has an exponential distribution (with mean  $N$ ) [2] and most lineages in a genealogical tree coalesce at early times. In contrast, the bulk of coalescence in a population under selection is significantly delayed - an effect corresponding to the comb-like appearance of the trees.

The asymmetry of branching is quantified in Fig. 3B which presents the distribution of weights at the level just below the MRCA. The strong bias toward extreme values of  $w$  in populations under selection is to be contrasted with  $w$ -independent distribution predicted and observed in the neutral case (see SI).

#### B. Correlation between ancestral weight and offspring fitness

Let us consider the whole population and trace the surviving lineages back in time, identifying all ancestors of the present day population  $t$ -generations in the past.

Figure 4A shows the distribution of the ancestral fitness (relative to the mean for that generation) at several time points in the past. This distribution becomes progressively shifted towards higher fitness as compared to the distribution for the whole population[22]. In the limit of large times, this distribution follows the fitness dependence of the non-extinction probability of a lineage [30, 31].

We shall be particularly interested in the time in the past when there is still a large number of ancestors, e.g. about  $a = 10^3$  at  $t = 100$ . Fig. 4B shows the scatter plot of the weight of ancestors versus their fitness advantage. Note that, by collapsing the points on the fitness axis, one gets the histogram shown in Fig. 4A for  $t = 100$ . We observe a strong positive correlation between the weight and the fitness of an ancestor. Higher fitness individuals in the past generations are not only more likely to survive, but they also leave more offspring conditional on the survival. Thus the weight of the ancestor, which can be determined from a reconstructed genealogical tree, can be used as a proxy for ancestral fitness: a quantity that one does not expect to know directly, except in the case of computer simulations! In SI, we provide plots of average ancestral fitness conditioned on its weight for various time points and parameter sets and confirm that the positive correlation between the weight and the fitness of ancestors holds quite generally. This correlation decreases as the time shifts further into the past.

Next we examine the correlation between the weight of the ancestor and the fitness of its surviving progeny. Consider a sample of genomes with size  $n$  and the corresponding genealogical tree. One expects genomes which are derived from relatively high fitness ancestors to belong to higher fitness classes in the present time. Since ancestral fitness correlates with weight, we expect higher weight ancestors to produce, on the average, higher fitness offsprings. Hence, given a genealogy of a sample of genomes, we examine the fitness  $\{f_1, \dots, f_{w_i}\}$  of the  $w_i$  offsprings in the sample stemming from the  $i$ -th ancestor that existed  $t$ -generations in the past. Let us define the mean,  $\bar{f}(w_i)$ , and the variance,  $\Sigma^2(w_i)$  over  $w_i$  offspring. In Fig. 4C and D, we show  $\bar{f}(w)/\sigma$  and  $\Sigma(w)/\sigma$  averaged over different population realizations at two different time points in the past for trees with sample size  $n = 100$  (see SI for other parameter sets). In both cases, the mean fitness of the derived genomes is an increasing function of the weight of their ancestor. From Fig. 4C, we also notice that the variance in the fitness of the derived genomes,  $\Sigma(w_i)$ , is higher for higher mutation rates. Consider a time closer to the root of a tree (e.g. right plot in Fig. 4C), where a lineage can carry a significant portion of the sample size. As expected, the value of  $\bar{f}(w)$  for such high-weight ancestors is close to zero (remember that  $f_i$  was defined relative to the population mean, so that the average of  $f_i$  over the whole sample is zero). At the same time  $\Sigma(w)/\sigma \rightarrow 1$  for ancestors with  $w$  approaching  $n$ . Interestingly, for the lineages which are still carrying a small weight at late in the coalescence

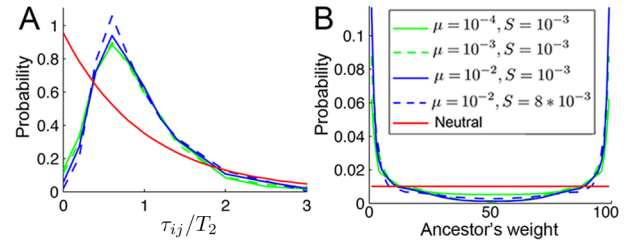


FIG. 3. Distortion in the shape of genealogies in the presence of selection. (A) Distribution of pairwise coalescent time, scaled with its mean,  $T_2$ . (B) Probability of an ancestor to carry weight  $w$  when there are  $a = 2$  lineages left in the genealogical tree of  $n = 100$  samples. Distributions based on 8000 random samples and population replicas.  $N = 64000$  and  $\epsilon = 0.1$  in both panels.

process, the value of  $\bar{f}(w)$  is clearly negative.

High-fitness genomes typically merge first in a tree and form high-weight ancestors. This fact is seen in the distribution of the pairwise coalescent time,  $\tau_{ij}$ , shown in Fig. 3A. Averaging  $\tau_{ij}$  over all  $\{i, j\}$  pairs of genomes in a population gives the mean coalescent time  $T_2$ . Now, consider the average of  $\tau_{ij}$  conditioned on the fitness of the two genomes: Fig. 4D shows a heat map of  $\bar{\tau}(f_i, f_j)/T_2$ . For two genomes both with high-fitness, the average coalescent time is shorter than  $T_2$ . This is because they are likely to be relatively recent lineages emanating from the “nose” of the distribution [20]. This observation is the key to the proposed fitness inference method.

### C. Relative fitness inference based on the reconstructed genealogy

Above we have reviewed the different ways in which the shape of the genealogical tree of the population under selection differs from a neutral one and have demonstrated the correlation between ancestral weights and the fitness of offsprings. We now show that this insight can be converted into a method for inferring relative fitness of genomes within the sample.

To that end, let us consider a randomly chosen set of  $n$  genomes from a population and use standard phylogenetic tree-building methods (see SI) to approximately reconstruct the genealogy of the sample. The accuracy of the reconstructed genealogy compared to the actual genealogy, known exactly from the forward simulation of population dynamics, is discussed in the SI. It increases with the neutral mutation rate  $\mu_0$ : in the biologically plausible regime of  $\mu_0/\mu \approx 10$  considered here, it proves more than adequate to enable meaningful inference.

Next, based on the reconstructed tree, we associate with each leaf  $i = 1, \dots, n$  a *fitness-proxy score* (FPS),  $\phi_i$ , defined by its lineage within the tree. Specifically, we



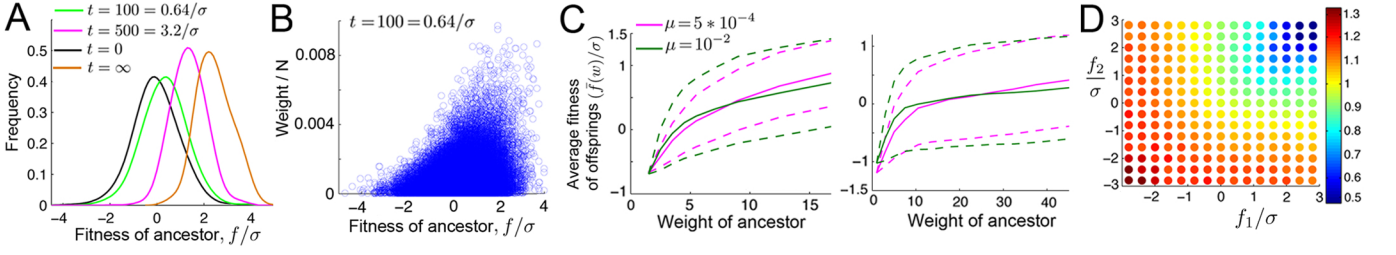


FIG. 4. Correlation between fitness, weight and coalescent time. (A) Fitness distribution of the ancestors of the whole population, for a few time intervals in the past. Fitness is measured at the time the ancestor existed. Shown in the legend is also the time values in the unit of  $1/\sigma$ .  $N = 64000$ ,  $\epsilon = 0.1$  and  $s = 2 \times 10^{-3}$  in all the panels.  $\mu = 10^{-3}$  in (A), (B) and (D). (B) Scatter plot of weight versus the fitness of the ancestors at  $t = 100$  generations ago. (C) Average fitness of offsprings as a function of the ancestral weight in a sample of size  $n = 100$  at two different time slices in the past shown using solid lines. The dashed lines represents the standard deviation,  $\Sigma(w)/\sigma$ , above and below the mean,  $\bar{f}(w)/\sigma$ . The two time points are chosen to be the first time that the tree carries a lineage with weight greater than 15% and 40% of the sample size, respectively. Since the curves for various parameter sets were similar, for the sake of clarity, we only show them for two sets. (D) Heat map of mean pairwise coalescent time as a function of the fitness of the two involved genomes,  $f_1/\sigma$  and  $f_2/\sigma$ , normalized by the mean pairwise coalescent time for the whole population:  $\bar{\tau}(f_1, f_2)/T_2$ .

define  $\phi_i$  as a linear discriminator in the form

$$\phi_i = \sum_{k=1}^{m_i} \Theta(t_{a_k(i)}/T_2)[w_{a_k(i)} - w_{a_{k-1}(i)}] \quad (1)$$

where  $\{a_k(i)\}$  is the lineage of genome  $i$ , starting with the genome itself as  $a_0(i)$  and running the length,  $m_i$ , of the lineage (i.e. the number of nodes) until the root of the tree. When an ancestral lineage  $a_{k-1}$  merges with an internal node  $k$ , it forms a new ancestral lineage  $a_k$ . The time of formation of the corresponding internal node is denoted by  $t_{a_k(i)}$ . The parameter  $T_2$  is the estimate of the average pairwise coalescent time, obtained from the sampled genomes. Finally,  $\Theta(x)$  is a "soft step" function (a.k.a. Fermi function):  $\Theta(x) = (1 + \exp(\beta(x/x_* - 1)))^{-1}$  parametrized by the position of the step  $x_*$  and its characteristic width  $\beta$ . If the  $\beta \gg 1$  function  $\Theta(x)$  steps abruptly from one to zero as  $x > x_*$ , so that  $\phi_i = w_{a_*} - 1$  where  $a_*$  is the oldest ancestor in the lineage with  $t_{a_*} < x_* T_2$ . For  $\beta \sim 1$  the FPS is defined by a weighted sum of ancestral weights from the  $t_a \sim x_* T_2$  "era" (see SI for details).

The logic behind our heuristic choice of the specific form of  $\phi_i$  is to exploit the correlation between the offspring fitness and ancestral weights, which, at least on the high fitness/ high weight end of the distribution, decreases for  $t_a > T_2$ , because at long times even the lineages originating from high fitness ancestors spread all over the surviving population. Hence we choose  $x_* < 1$ : specifically the results below were obtained with  $x_* = 0.5$  and  $\beta = 5$ , but in the SI we examine the performance of the ranking algorithm as a function of the parameters and demonstrate that nearly optimal performance (at least for the present form of the FPS) is achieved for a broad range of  $x_*, \beta$ . Critically, normalization of  $t_a$  to the characteristic time of coalescence for the sample,  $T_2$ , essentially eliminates the need to know the evolutionary parameters of the population, such as its effective  $\mu/s$  or

$N$ .

We rank genomes according to their  $\phi_i$  score and compare this ranking with the actual fitness of each genome. In addition to inferring relative fitness, it is useful to know how genetically close a given genome is to the fittest in the sample. Hence, for each genome we define  $d_i$  as average of its Hamming distance to the fittest 10% genomes in the sample. Fig. 5A-B shows the results of the ranking for two  $n = 200$  samples from the populations that already appeared in Fig. 2A and B. We observe a strong correlation between FPS ranking and the actual fitness in general and the "tendency" (quantified below) for the fittest genomes of the sample to show up in the top ranks. In addition, highly ranked genomes have smaller  $d_i$  values indicating that they are genetically close to the fittest 10%.

The above observations are confirmed and quantified by the statistical data obtained by repeating the comparison for 8000 independent population samples and different sets of parameters. Specifically, Fig. 5C shows mean fitness conditional on the FPS ranking and Fig. 5D shows the mean rank conditional on actual fitness (normalized by  $\sigma$ ) for two different values of  $\mu$ . Fig. 5E shows mean distance from the fittest conditional on the FPS ranking (for four different values of  $\mu$ ), with distance normalized to  $\Delta_{10\%}$  defined as the average  $d_i$  amongst the fittest 10%. Remarkably, we observe that  $d/\Delta_{10\%}$  for the highest ranked genomes gets close to one, indicating good convergence, in the sense of genetic distance, of the top ranked genomes to the fittest set. Further analysis of the algorithm's performance, as well as additional parameter sets including the case of  $\epsilon = 0$ , can be found in SI. For example, the top 10% ranked genomes fall into the top 20% fittest with probability  $0.7 - 0.9$ , depending on parameters, with more accurate inference achieved at lower mutation rates.

Looking at Fig. 5A we note greater dispersion in the fitness of the highly ranked subset, then in distance. In-

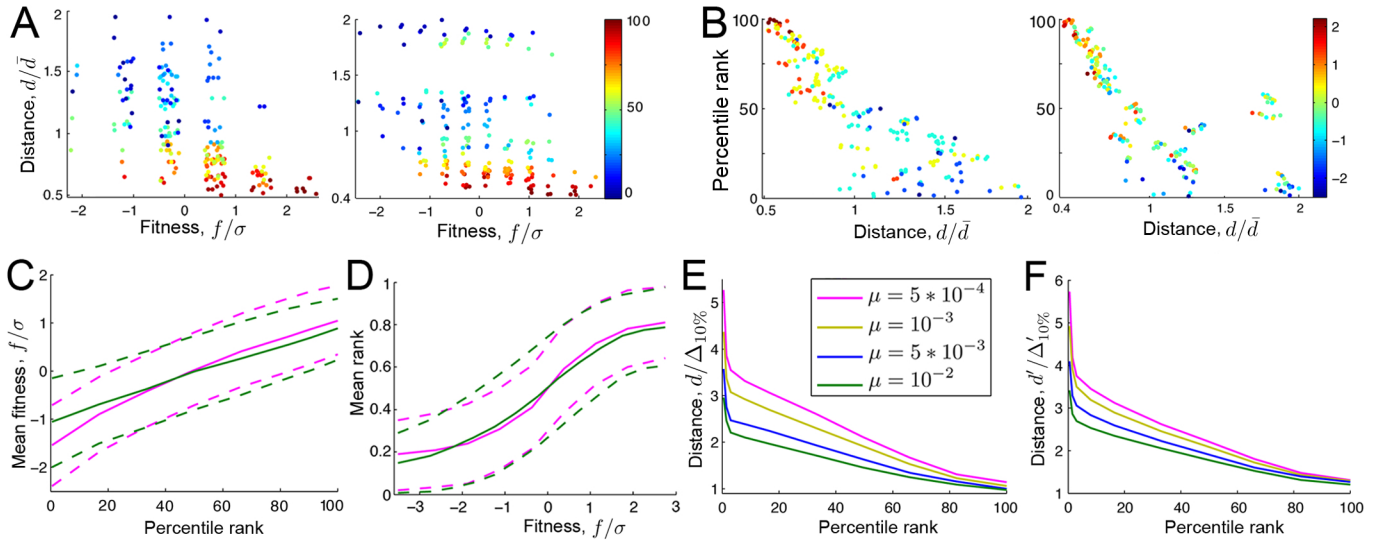


FIG. 5. Performance of the ranking algorithm. (A) Heat map of rank as a function of fitness and average distance to the fittest 10% genomes. Distance  $d$  is normalized by its mean  $\bar{d}$ . Left and right panels correspond to two samples of size  $n = 200$  drawn from the same populations as Fig. 2A ( $\mu = 10^{-3}$ ) and Fig. 2B ( $\mu = 10^{-2}$ ), respectively. (B) Scatter plot of rank versus distance to the top 10% fittest genomes (colormap represents  $f/\sigma$ ). The panels correspond to the same trees as in (A). (C) *Solid lines*: mean fitness as a function of rank. *Dashed lines*: standard deviation above and below the mean ( $\mu = 5 \times 10^{-4}$  and  $10^{-2}$ ). (D) Same as (C) for mean rank as a function of fitness. (E) Mean genetic distance to the top 10% fitness set, normalized by  $\Delta_{10\%}$  (see text) as a function of rank. (F) Mean genetic distance to ancestors of the generation at one turnover time in the future, normalized by  $\Delta'_{10\%}$  (see text) as a function of rank.  $N = 64000$ ,  $\epsilon = 0.1$  and  $s = 2 \times 10^{-3}$  in all cases.

deed, some genomes which are not among the fittest still can be genetically close to the fittest subset: e.g. note in Fig. 2A-B the genomes with blue color located close to the mostly orange/red clusters on the right side of the trees. This is because genetic distance is dominated by neutral mutations  $\mu_0 \gg \mu$ , and is less susceptible to fluctuations than fitness, which is defined by a much smaller number of non-neutral mutations. To the extent that genetic relatedness is defined by the distance, the latter is essential for identifying within the sample the closest relatives of future populations. Taking advantage of ready accessibility of evolutionary future within our simulations, we have directly tested the ability of our approach to identify within samples, the genotypes that are close to future populations. For each sampled genome, we define  $d'_i$  as the average of its Hamming distance to all of the genomes in the current population that are direct ancestors of the population in a generation about one genetic turnover time in the future (we know these ancestors from the forward simulation). Typically, less than 100 individuals from the current population of  $N = 64000$  have descendants in this future population. In each case we normalized the distances by  $\Delta'_{10\%}$  defined as the average of the smallest 10% values of  $d'_i$ . Fig. 5F shows  $d'_i/\Delta'_{10\%}$  conditional on the FPS ranking. We again observe that  $d'/\Delta'_{10\%}$  for the highest ranked genomes gets close to one, indicating that the top ranked genomes are indeed the closest to the ancestor of future generations. This means that FPS ranking makes possible to identify common genetic elements that future

populations inherent from the present one.

In summary, above results clearly indicate the power of the proposed inference method. The performance of the method improves monotonically with the increasing sample size (see SI): it degrades significantly, compared to the results presented above, for  $n < 100$  but approaches saturation for  $n > 200$ .

#### IV. DISCUSSION

Whereas one often thinks of evolution occurring on geological time scales, evolutionary dynamics can also unfold swiftly as it does in bacteria acquiring antibiotic resistance, in HIV evading CTL response in the course of infection or in the progression of an aggressive cancer. Recent advances of sequencing [32, 33] have made it possible to extensively sample such rapidly evolving populations. The amount and quality of genomic data on populations will only continue to increase, accentuating the challenge of extracting more information from sampled genomes. Here, we have demonstrated that the shape of genealogical trees contains much more information than merely the evidence for (or against) selection within population. As a proof-of-principle we have formulated a method for estimating relative fitness of individual genomes sampled from a fitness diverse but otherwise unstructured population, in the absence of any information other than genomic sequence. This provides the possibility of forecasting the common genotype of the

future on the time scale of genetic turnover.

Our demonstration was based on a vast simplification of biological and ecological reality. Our model assumed fixed population size and constant environment; it neglected epistasis and assumed all non-neutral mutations (both deleterious and beneficial) to have the same effect size. While we have, within the model considered, explored a biologically interesting range of parameters, it would be useful to extend the study to a broader class of models. Yet, we expect the proposed method to be quite robust, because it is based on the very fundamental aspect of evolutionary dynamics, realized when population and the mutation rate are sufficiently large to harbor substantial non-neutral diversity, and when fitness differentials between individuals are formed by the contributions of numerous weakly selected loci rather than a small number of strong ones. In this multi-locus weak selection regime, surviving lineages in the course of time move from the nose of the fitness distribution towards the center, in an biased diffusion fashion. The correlation between early coalescence and rapid increase of ancestral weight along the lineages with high relative fitness, derives from the continuous genetic turnover of the population described above. This turnover occurs in traveling waves models corresponding to the continuous adaptation scenario [18, 19] and in the dynamic mutation-selection balance [25] which involves both deleterious and compensating beneficial mutations (as well as in the case of pure purifying selection [23] ( $\epsilon = 0$ ) provided that  $\mu/s$  is large enough to make Muller's ratchet click fast [34].

A detailed statistical analysis of the way lineages propagate along the fitness axis could allow to improve FPS by optimizing the tradeoff between gaining more information about a particular lineage by tracking it further back in time and the loss of predictive power due to the fact that beyond the genetic turnover time even lineages of the fittest ancestors spread all over the fitness distribution. Presently we have dealt with the problem heuristically by focusing on the coalescence sequence for each lineage up to about  $0.5T_2$ . The advantage of our simple heuristic approach is that it is more likely to be model independent than the more fine-tuned methods.

It would be interesting to extend the fitness inference method to recombining populations. This should be relatively straight-forward as long as genetic turnover time is faster compared to the inverse recombination rate. For a chromosome with an approximately uniform crossover probability, this condition defines a characteristic length below which loci coalesce in essentially recombination-free genealogies. Roughly, the asexual coalescent considerations would apply to a 1cM size locus provided that it harbors  $\sigma > 10^{-2}$ . More careful analysis is however necessary in order to deal with the Hill-Robertson effect or *genetic draft* [35, 36] caused by the transient linkage of the locus to the rest of the genome which effectively adds noise, reducing effectiveness of selection on the individual loci.

Clearly, the highest priority for the future would to

test the method on experimental or epidemiological data. Applications are possible wherever genomic data is available for fitness-diverse, but otherwise unstructured populations. Genomic data from single cell sequencing of tumors [33] or from localized influenza outbreaks [37] are among the interested possibilities to be considered. For example, it would be interesting to compare the proposed method with the clustering-based approach of [38] to predicting antigenic evolution of influenza A. In addition to predicting which genotypes are more likely to appear in future generations, fitness inference method could be used for QTL mapping [39] with FPS-based ranking being the quantitative phenotype that could be used to identify highly adaptive or deleterious alleles.

## V. ACKNOWLEDGEMENTS

We thank Richard Neher, Daniel Balick and Sidhartha Goyal for many useful discussions. AD was supported by HFSP RFG0045/2010 and NSF PHY11-25915 while BIS acknowledges support of NIGMS R01 GM086793.

## APPENDIX A: SUPPLEMENTAL INFORMATION

### A. Evolutionary simulations

The simulations are done using a custom written Python code, available upon request. The evolution is based on a discrete time Wright-Fisher model with population size  $N$ . Each generation  $t$  undergoes separate selection and mutation steps. To implement selection, each individual  $i$  produces a Poisson-distributed number of gametes in the next generation with parameter  $\exp(f_i - \alpha)$ . Here  $f_i = F_i - \bar{F}$  is the fitness advantage of individual  $i$  relative to the mean fitness of the population  $\bar{F}$ , and  $\alpha = \frac{N(t) - N}{N}$  ensures an approximately constant population size around  $N$ . Individual genomes are defined as binary strings,  $g_k$  with  $k = 1, \dots, L$  and the number of loci,  $L = 10^5$ , chosen large enough to exceed the number of segregating polymorphisms in the simulated population. Consistent with infinite site approximation, new mutations flip the  $g_a$  binary value from zero to one.

At each generation, the beneficial and deleterious mutations arise with probability  $\epsilon\mu$  and  $(1 - \epsilon)\mu$  and have a fitness effect of  $\pm s$ , respectively. We also record the forward genealogies during the simulations. The above process is repeated for a specified number of generations. Various information on the dynamics of the evolution are measured after an equilibration time to remove transient effects from the initial conditions. In the parameter regimes studied, we found that  $10^4$  generations was generally sufficient.

Given that we perform forward simulations and keep track of the genealogies, the simulations are computationally intensive. Therefore, the maximum population

size that we simulated was  $N = 64000$ . The mutation rate was varied from  $\mu = 10^{-4}$  to  $\mu = 10^{-2}$  and the selection coefficient from  $s = 10^{-3}$  to  $\mu = 8 * 10^{-3}$ . For the parameter combination where  $N = 64000$ ,  $\epsilon = 0.1$  and  $\mu = 10^{-4}$  (beneficial mutation rate  $10^{-5}$  and deleterious mutation rate  $9 * 10^{-5}$ ), only a couple of clones are segregating in the population (see below). This parameter combination serves as the boundary between the multi-site selection regime and the selective sweep regime. For smaller mutation rates, given that  $N = 64000$ , the population is monoclonal and enters the regime of selective sweeps.

Below, we present some results on the clonal diversity, as well as the speed of adaptation, for various parameters that we have simulated. In Fig. 2A and B of the main text, we showed two examples of fitness distribution. In Fig. 6, we show some more examples. Assume there are  $c$  clones in the population, with sizes  $n_1, \dots, n_c$ . Note that  $\sum_{i=1}^c n_i = N$ . To see how many clones with significant size are segregating, we can define the participation fraction:  $Y = \langle \sum_{i=1}^c (\frac{n_i}{N})^2 \rangle$ . This quantity is equal to the probability that two randomly chosen genomes belong to the same clone. Fig. 7A shows the participation fraction for various sets of parameters. For  $N = 64000$  and  $\mu = 10^{-4}$ ,  $Y \approx 0.4$ , which means that at low  $\mu$ , there is a significant probability that two randomly chosen genomes come from the same clone.

In the regime of our interest where many mutations simultaneously segregate, it is well known that the competition between mutations slows down the rate of the adaptation (Hill-Robertson or Fisher-Muller effect) [20, 21]. In Fig. 7B, we present the speed of the adaptation, i.e. the rate of change of the mean fitness, normalized by its expected value in the selective sweep regime. In the later regime, the beneficial mutations are rare enough that only a single mutation segregates at a time, and assuming the deleterious mutations are purged, the expected speed of adaptation is  $v = 2N\mu\epsilon s^2$ . As we see in Fig. 7B, for the parameter combination  $N = 64000$  and  $\mu = 10^{-4}$ , the adaptation rate is only around a quarter of its expected value in the selective sweep regime.

## B. Tree reconstruction

In the first step, we use the neighbor-joining algorithm [40, 41] to reconstruct the tree topology. The input distance matrix for this algorithm is simply given by the pairwise difference of sequences (Hamming distance) including both neutral and non-neutral mutations. The time to the common ancestor of two individuals is proportional to the number of neutral genetic differences between them. For a real data set, one may use a more realistic substitution model to infer the divergence time between pairs of genomes. We have considered the values of the non-neutral mutation rate over a few orders

of magnitude. The neutral mutation rate was always set to 10 times the value of the non-neutral mutation rate. We use the neighbor-joining algorithm only to infer the topology of the tree. We do not use the length of the edges that are calculated in this algorithm. The reason is that we want all the leaves of the tree to be located at the current time and have the same distance to the root.

In the next step we find the root of the tree based on the parsimony method. Each point on the tree divides the sample into two groups. The root should be located at a point where the similarity between the two groups is minimal. We count the number of mutations which exist in both groups and assign the root to a point where this number is minimal.

In the last step, we assign the height (time interval to the present time) of each node in the tree. The lengths are calculated as in the UPGMA algorithm [41]. In this algorithm, the total branch length from a tip down to any node is half of the average of the distance between all the pairs of genomes whose most recent common ancestor is that node. We consider a node only after all the nodes below it have their heights assigned. We start from the bottom, namely, the nodes which are connected to two leaves. The height of these nodes are calculated similar to the UPGMA algorithm: the height is equal to the half of the mutational distance between the pair of the genomes below that node. For other internal nodes, we also calculate the putative height as half of the distance between all the pairs whose most recent common ancestor is that node. The height of the node is the maximum between this putative distance and the height of all the internal nodes below the considered node.

We evaluated the performance of the above tree reconstruction algorithm in all different parameter ranges by comparing the reconstructed tree with the actual genealogy. In all the cases, the performance was satisfactory. In Fig. 8, we show examples of the performance of the above algorithm for four different mutation rates. For each rate, a sample of size  $n = 100$  is selected. In Fig. 8A, we show the sequence distance for all the  $(n-1)n/2$  pairs in the sample versus the real divergence time. These distances are the input of the above algorithm. In Fig. 8B, we show the reconstructed divergence time (inferred from the reconstructed tree) for all the pairs. The validity of the above algorithm is reflected in the fact that the relation between these two times seems to be linear. The slope of the line is irrelevant, since, it only reflects an scaling factor, i.e. the estimation of the mutation rate. As we see in Fig. 8, for higher mutation rates (e.g.  $\mu = 5 * 10^{-3}$  and  $\mu = 10^{-2}$ ) where there is tens-hundreds of differences between a typical pair of genomes, the assumption of the neutral mutation rate being 10 time that of  $\mu$  is unnecessary. In these cases, there is enough diversity that even setting the neutral mutation rate equal to  $\mu$  would be sufficient.

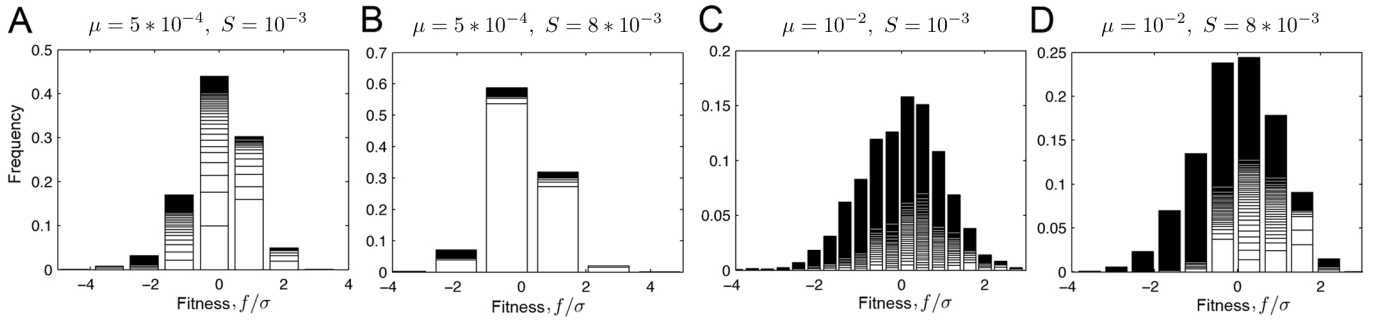


FIG. 6. Fitness distribution at one time slice. The mutation rate and selection coefficient for each case is written on the top of the corresponding panel.  $N = 64000$  and  $\epsilon = 0.1$  for all the panels. Each bin corresponds to a fitness class and each class is composed of several clones. The height of each box within each bin represents the size of a clone. Larger clones are stacked on the bottom. The dark band on top of each bin correspond to small clones.

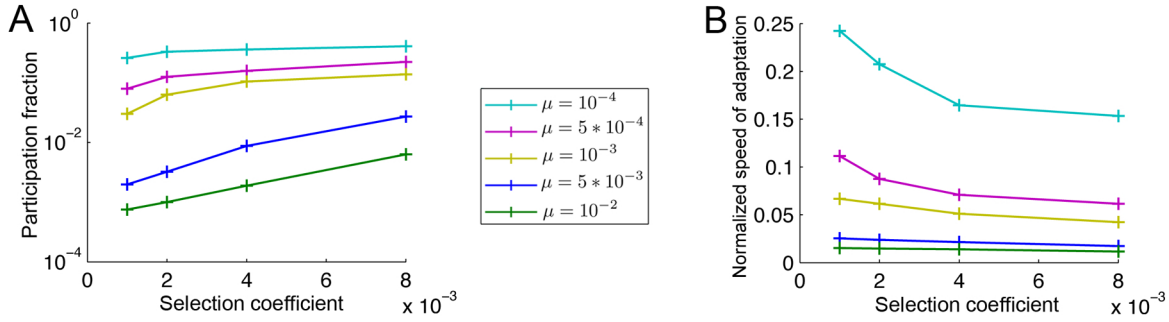


FIG. 7. Clonal structure and adaptation rate. (A) The participation fraction,  $Y$ , for different parameter values.  $Y$  is the probability that two randomly chosen genomes belong to the same clone. For all the curves,  $N = 64000$  and  $\epsilon = 0.1$ . When  $\mu = 10^{-4}$  (beneficial mutation rate  $10^{-5}$  and deleterious mutation rate  $9 * 10^{-5}$ ), the values of  $Y$  become significant ( $> 0.1$ ). This implies that the dynamics is at the boundary between the multisite selection regime and the selective sweep regime. (B) Speed of adaptation, normalized by its expectation value in the limit of selective sweep  $2N\mu\epsilon s^2$ .

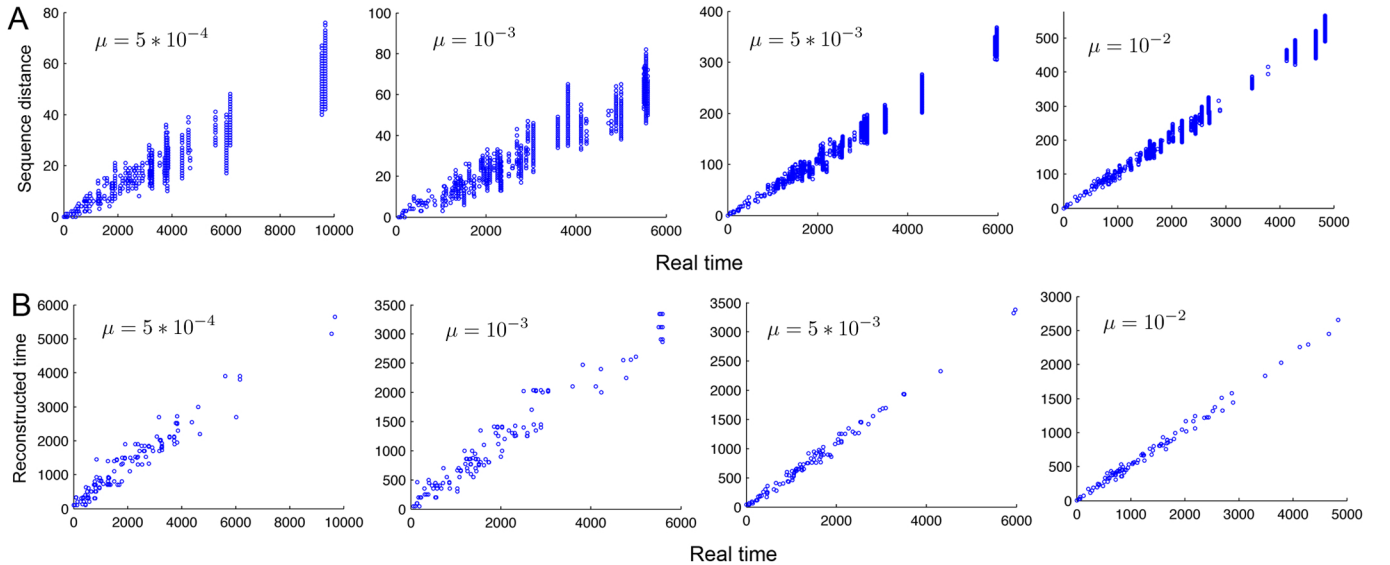


FIG. 8. Examples of tree reconstruction for sample size of  $n = 100$ .  $N = 64000$ ,  $\epsilon = 0.1$  and  $s = 2 * 10^{-3}$  in all cases. (A) Scatter plot of the Hamming distance between sequences versus the real divergence time for all the pairs in the sample. The non-neutral mutation rate for each case is shown in the associated plot. The neutral mutation rate was set to 10 times the value of the non-neutral mutation rate. (B) Scatter plot of the reconstructed divergence time between sequences versus the real divergence time for all the pairs in the sample. Each plot is associated to the same tree as in panel (A) for the same mutation rate.



### C. Weight Distribution

Consider a sample of size  $n$  and the corresponding phylogenetic tree. Assume looking at the tree at the stage where there are  $a$  lineages left. The ancestor  $i$  will carry a weight  $w_i$  where  $i = 1, \dots, a$  and  $\sum_{i=1}^a w_i = n$ . The values that  $w_i$  can take is anything between 1 and  $n - a + 1$ . For example, when there are only 2 ancestral lineages,  $w_i$  can be between 1 and  $n - 1$ . The statistics of the phylogenetic trees for neutral evolution are given by the Kingman's coalescent [3, 42]. In particular, the probability distribution of  $w_i$  is given by [43]:

$$P_{neu}(w_i|a, n) = \binom{n - w_i - 1}{a - 2} / \binom{n - 1}{a - 1} \quad (2)$$

For example, when there is only  $a = 2$  ancestors left in the tree, we get  $P_{neu}(w|2, n) = \frac{1}{n-1}$ , which is independent of  $w$ . In other words, when there are two ancestors left, each one can carry any weight between 1 to  $n - 1$  equally likely. The above formula can be derived solely based on the fact that, as one goes up in the tree, at each stage, any lineage is equally likely to coalesce with any other lineage regardless of the weight they are carrying or any other previous events in the tree. In the presence of selection, this is no longer the case and not all lineages are equally likely to coalesce. The probability of the coalescent between two lineages will depend on the history of previous merging events.

### D. Distortion in shape of genealogical trees

Here, we consider some quantities which reflect the differences between the shape of trees from non-neutral and neutral evolution. While inspecting trees in Fig. 2D and E, we notice that in the presence of selection it is more common for a leaf (sampled genome) to be connected to a long edge. In other words, it takes a long time for some leaves to merge to other lineages in the tree. Moreover, such leaves are more likely to belong to lower fitness classes, represented by blue and grey colors. In addition, number of lineages left in a tree as a function of time seems to be different. Here, we explore such points in more details.

At each instant of the time in a tree, one can consider what fraction of the remaining lineages are singletons. Singletons are defined as lineages with weight  $w=1$ . In Fig. 9A shows the average value of this fraction as a function of time. These curves are obtained by averaging over random samples and over population replicas. The time for each tree is linearly rescaled so that the current time is at 0 and the root is at 1. At time 0, all the lineages in a tree are singletons and the fraction is, therefore, one. At time 1, all the lineages have merged together and therefore no singleton lineage is left. As we see, the curve for the neutral case falls below the rest of the curves. One can ask, by looking at the number

of singletons as a function of time for a single tree, is it possible to tell whether or not this tree is from a neutral population? Later, we show that the separation between the neutral and non-neutral curves is large enough that one can differentiate with high confidence whether or not a single tree is neutral.

In the neutral case, the statistic of length of singleton edges was studied in [5] and is used in the Fu and Li's test for detecting departures from the Kingman's coalescent. As long as the sample size is not very small (e.g.  $> 5$ ), the expected total height of a neutral tree is nearly equal to  $2N$  where  $N$  is the population size. This quantity is almost independent of the sample size. The expected length of a singleton edge is given by  $\frac{2N}{n}$  where  $n$  is the sample size [5]. In our example of tree in Fig. 2 of the main text where  $n = 30$ , the average length of singleton edges is around  $1/30$  of the total height of the tree. This means that most of such edges should be much shorter compared to the total height of the corresponding tree. However, in the presence of selection, it is more common that some singletons survive even until close to the root of the tree.

One can also look at the fitness of the singleton lineage that is the latest to join the rest of the tree. Fig. 10 shows the fitness distribution using the same simulation parameters as in Fig. 2D. As we see, these lineages tend to belong to the unfit classes. This is a general pattern observed for all of the simulation parameters.

We have also considered the average number of lineages left in a tree as a function of time,  $\langle a \rangle_t$ . The result is presented in Fig. 9B. In the presence of selection, the number of lineages drops slower at early times compared to the neutral case. Under neutrality, when there are  $a$  lineages, the rate at which the next coalescent event happens is  $\binom{a}{2}/N$  [44]. This is the product of the coalescent rate between two random lineages,  $1/N$ , and the total number of pairs among  $a$  lineages,  $\binom{a}{2}$ . Therefore, the expected time a tree spends having  $a$  lineages is  $N/\binom{a}{2}$ . In this case, the coalescent events happen much faster on the bottom of the tree, where  $a$  is large, and most of the time in the tree is spent while having only a few lineages (also see Fig. 2F of the main text). On the other hand, in the presence of selection, coalescence times near the root are reduced compared to the total height of the tree. Alternatively, the external branches are longer compared to neutral expectations.

Under neutral evolution, since the coalescent events happen at rate  $\binom{a}{2}/N$ , when there are  $a$  lineages left, one has:  $\frac{d\langle a \rangle_t}{dt} \propto - \langle \binom{a}{2}/N \rangle$ . Therefore, the ratio  $-\frac{d\langle a \rangle_t}{dt} / \langle \binom{a}{2}/N \rangle$  which is the coalescent rate between two random lineages remains constant. Fig. 9C shows the coalescent rate normalized by its value at time  $t = 0$ . For the neutral case, this rate remain constant, as expected. However, in the presence of selection, the rate increases for further time back in the tree. The reason for this is that, for times further back in the tree, the ancestral lineages are more likely to have belonged to the leading edge of the fitness distribution at the time they

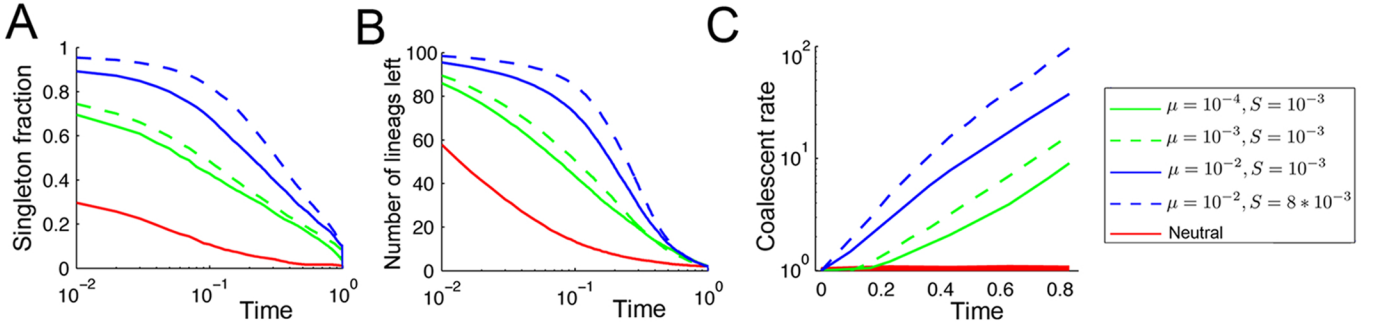


FIG. 9. Statistics on lineages in phylogenetic trees. For all the panels, the sample size is  $n = 100$ ,  $N = 64000$  and  $\epsilon = 0.1$ . (A) Average fraction of singleton lineages left in a tree as a function of time. The time for each tree has been linearly rescaled so that the root is at  $t = 1$  and the current time is 0. (B) Average number of lineages left in a tree as a function of time. The time has been linearly rescaled as in part (A). (C) Coalescent rate between two random lineages as a function of time. The rate is normalized by its value at time  $t = 0$ . The ‘effective population size’,  $N_e$ , would be defined to be inversely proportional to coalescent rate.

existed (see Fig. 4A of the main text). Therefore, they coalesce at a faster rate compared to the bottom of the tree where lineages are spread over the fitness distribution [22].

Increase in the coalescent rate is sometimes interpreted as a reduction in the effective population size (denoted by  $N_e$ ). However, not all aspect of the coalescent process under selection, such as the weight distribution or fraction of singleton lineages, can be accounted for by only introducing an effective population size [11]. This fact also manifests itself in the distribution of polymorphisms in a sample of genomes. Under neutrality (in the limit of infinite-site model), the probability that a derived allele appears in  $w$  individuals out of  $n$  sampled genomes is proportional to  $1/w$ . This behavior is a consequence of both the weight distribution and the length of coalescent intervals. To see this, note that in order to appear in  $w$  genomes in the sample, a mutation must have occurred on an ancestor with weight  $w$ . Assume this ancestor existed when there was  $a$  lineages in the tree. The probability that 1 of the  $a$  ancestors carried weight  $w$  is  $a * P_{neu}(w|a, n)$ . The average time a tree spends having  $a$  lineages is proportional to  $1/\binom{a}{2}$ . Summing over all possible  $a$ ’s gives:  $\sum_{a=2}^n a P_{neu}(w|a, n) \frac{1}{\binom{a}{2}} = \frac{2}{w}$ , which is the usual one over frequency dependence. In Fig. 11 we show the frequency distribution of neutral polymorphisms in the presence of selection. The distribution first drops more like  $1/w^2$  for small frequencies and then bends upward for higher frequencies where  $w > n/2$  [36].

#### E. Correlation between weight and fitness of ancestors

In Fig. 4A of the main text, we showed the distribution of the fitness of the ancestors for certain time intervals in the past for a set of parameters. Let us denote this distribution by  $\phi_t(f)$ . In the limit of large times, this distribution is equal to the fitness distribution for the

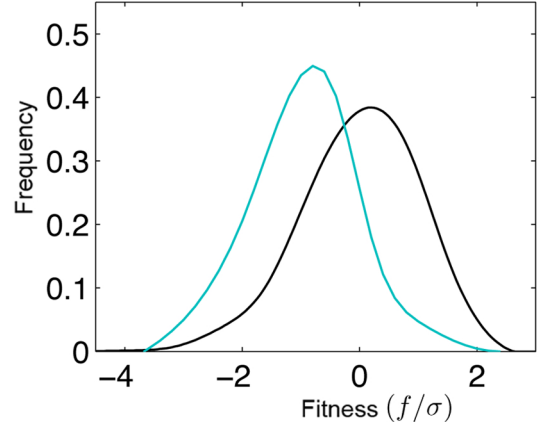


FIG. 10. Fitness distribution of the singleton lineage which connects to the tree the latest is shown in cyan color. The black curve presents the fitness distribution of the whole population which is the same as fitness distribution of the sampled genomes. The distributions are obtained by averaging over random samples and over population replicas.  $N = 64000$ ,  $\mu = 10^{-3}$ ,  $s = 2 * 10^{-3}$  and  $\epsilon = 0.1$ .

common ancestor of the whole population,  $\phi_\infty(f)$ . In Fig. 4B, we also showed the scatter plot between the weight of ancestors and their fitness advantage for  $t = 100$  generations in the past. The scatter plot represents the joint distribution of weight and fitness of ancestors,  $\phi_t(f, w)$ .

One can consider the expected fitness of an ancestor given its weight,  $\bar{f}_{anc}(w, t) = \sum_f f * \phi_t(f|w)$ . Fig. 12A shows  $\bar{f}_{anc}(w, t)/\sigma$  as a function of  $w/N$  for  $t = 100$  and  $t = 500$  in log-log scale. The dependence seems to be linear, namely,  $\bar{f}_{anc}(w, t) \propto w^{m(t)}$ , where  $m(t)$  is the slope of the lines in Fig. 12A. This slope depends on the time, and of course, other parameters such as  $N, \mu$ , etc. Fig. 12B shows  $m(t)$  as a function of time for different sets of parameters. For each set of parameters, the time

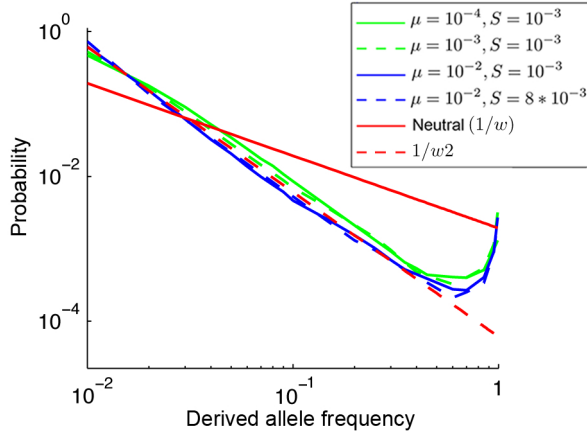


FIG. 11. Distribution of neutral polymorphism in a sample of size  $n = 100$ . The dashed red line shows the probability distribution which depends on  $1/w^2$ , as opposed to  $1/w$  (neutral case).  $N = 64000$ ,  $\epsilon = 0.1$ .

axis has been rescaled with the fitness variance for the corresponding parameter set,  $\sigma(N, \mu, \epsilon, s)$ . As we see, the slope  $m(t)$  drops as a function of time. In other words, the correlation between the weight and the fitness of ancestors reduces as one goes further back in time.

In the main text, we also presented some results on the relation between the weight of an ancestor in a tree,  $w_i$ , and the fitness of the  $w_i$ 's genomes in the sample which are derived from that ancestor. In particular, we focused on the mean,  $\frac{1}{w_i} \sum_{j=1}^{w_i} f_j$ , and the vari-

ance,  $\frac{1}{w_i} \sum_{j=1}^{w_i} (f_j - \bar{f}(w_i))^2$ . The average of these quantities over random samples of genomes and over population replicas are denoted by  $\bar{f}(w_i) = \langle \frac{1}{w_i} \sum_{j=1}^{w_i} f_j \rangle$  and

$\Sigma^2(w_i) = \langle \frac{1}{w_i} \sum_{j=1}^{w_i} (f_j - \bar{f}(w_i))^2 \rangle$ . In the main text, we

only presented these quantities for two parameter sets. In Fig. 13A and B, we show  $\bar{f}(w)/\sigma$  and  $\Sigma(w)/\sigma$  for more parameter sets. The sample size is  $n = 100$  and the results are shown for two different time points. One of the time points is chosen to be the first time that the tree carries a lineage with weight greater than 15% of the sample size. The other time point corresponds to the first time the weight of a single lineage becomes greater than 40% of the sample size.

## F. Fitness proxy score

Consider a sample of  $n$  genomes and the corresponding reconstructed phylogenetic tree. Although there is always a positive correlation between the weight of an ancestor and its fitness and the fitness of its derived genomes, both of these correlations drop as one goes further back in time. When most of lineages have condensed

into high-weight ancestors, the average fitness of the offsprings of such ancestors is close to zero and there is little correlation between the weight and the fitness of the offsprings (see right plot in Fig. 4C of the main text). The variance in the fitness of the offsprings also becomes close to the population fitness variance  $\sigma$ . In other terms, all of the derived genomes of such high-weight ancestors are, more or less, evenly distributed across the fitness distribution.

This is consistent with our observations in Fig. 4D of the main text. As the coalescent time for a pair of genomes increase, the difference in the fitness of the two genomes increases as well. This means, as  $\tau_{ij}$  becomes larger compared to  $T_2$  (the region covered in yellow and red colors in Fig. 4D), there is less information about the fitness of the pair of genomes involved. For example, one can have high fitness and the other one low fitness, or both can have average fitness. In other words, when the coalescent time for a pair of genomes becomes larger compared to the population average  $T_2$ , there is more uncertainty on the fitness of that pair of genomes.

Because of the above argument, we do not want the scoring scheme to be affected by the coalescent events far back in the tree. In addition, as we saw in Fig. 4D of the main text, when the fitness of two genomes is higher, the coalescent time between them is shorter compared to the mean pairwise coalescent time for the whole population  $T_2$ . The correlation between weights and fitness is also stronger for earlier times. Therefore, the earlier a coalescent event, the more it should affect the scores. It is important to have a sense of 'early times' or 'late times' in a tree. We use the empirical value of the mean pairwise coalescent time (i.e. estimate of  $T_2$  from the sample) for this purpose. In the algorithm, the time values appear only in the form of ratios. So having a correct estimate of the mutation rate is irrelevant.

To incorporate the above ideas, we introduced a threshold time  $t^* = x_* \times T_2$  and have the coalescent events which happen at a time further back compared to  $t^*$  contribute progressively less on the score. On the other hand, the coalescent events earlier than this stage will be progressively more important in the scoring scheme. In order to do this, we introduced the function  $\Theta(t)$  with a Fermi-Dirac form, shown in Fig. 14. In the results shown in this paper on the performance of the algorithm, we set  $t^* = 0.5 \times T_2$ , where  $T_2$  is the average pairwise coalescent time. We checked the performance for various values of  $t^*$ . We found that, in general, the results are very robust within a range of  $0.4 \times T_2 < t^* < T_2$  (see below).

We will now study the performance of the fitness ranking algorithm from other angles. Let us examine the fitness of the genome with the highest rank. Consider the sorted vector  $F = [f_1, \dots, f_n]$  which contains the fitness values for all the sampled genomes. In Fig. 15A, we show the probability for the fitness of the top 10% ranked genome to belong to the top 50% of fitness values of sampled genomes. In all cases, this probability turns out to be around 0.9. We can also look at the

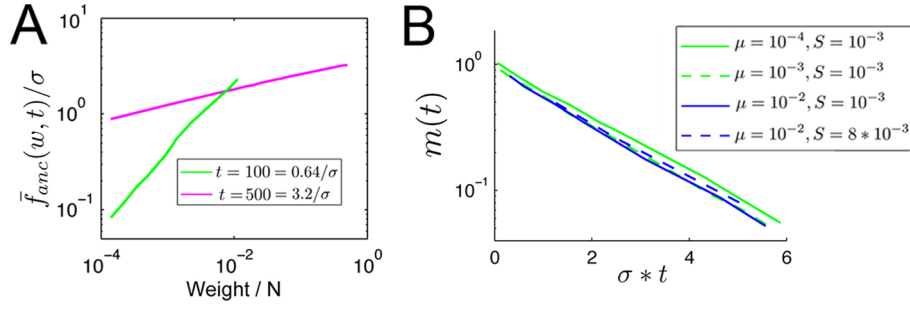


FIG. 12. Correlation between the weight and fitness of ancestors. (A) Average fitness of an ancestor conditional on its weight for two time intervals. Note the log-log scale.  $N = 64000$ ,  $\epsilon = 0.1$ ,  $\mu = 10^{-3}$  and  $s = 8 * 10^{-3}$ . (B) Fitting a line to the curve in part (A) gives a time dependent slope  $m(t) = \log(f_{anc}(w, t))/\log(w)$ . The slope  $m(t)$  is plotted as a function of time for a few different parameter values. Note the log scale on the y-axis. The time for each parameter set has been rescaled by the corresponding  $\sigma$ . Sample size  $n = 100$ ,  $N = 64000$  and  $\epsilon = 0.1$ .

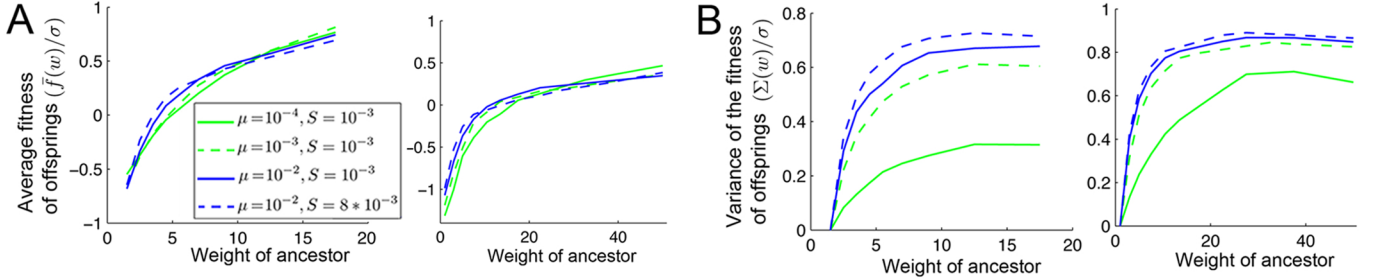


FIG. 13. Correlation between the weight and fitness of offsprings. (A) Average fitness of genomes as a function of the ancestral weight for two different time slices in the past. (B) Variance in the fitness of genomes as a function of the ancestral weight for two different time slices in the past.  $N = 64000$ ,  $\epsilon = 0.1$  in all cases.

probability for the fitness of the top ranked genomes to belong to the top 20% values in  $F$ . This probability is presented in Fig. 15B using solid lines. However, note that some of the sampled genomes can have equal fitness and, therefore,  $F$  contains duplicate values. As a consequence, the above probability for a randomly chosen genome is not necessarily  $1/5$ . To see this, consider the following explanatory example. If  $F$  contains only values  $-1$  and  $1$  in equal proportion, then the top 20% values in  $F$  is the same as the top 50%. In this case, a random genome will have a fitness belonging to top 20% values in  $F$  with probability  $1/2$ , instead of  $1/5$ . To take into account this fact, in Fig. 15B we show the above probability for a randomly chosen genome using dashed lines. For the highest mutation rate and the lowest selection coefficient,  $\mu = 10^{-2}$  and  $s = 10^{-3}$ , the probability for a randomly chosen genome is indeed close to the value of  $1/5$ . Lastly, in Fig. 15C, we present the probability for the fitness of the top ranked genomes to belong to the top 10% values in  $F$ . Again, we use the dashed lines to show this probability for a randomly chosen genome.

We have checked the performance for various values of  $t^*$ . We found that, in general, the results are very robust within a range of  $0.4 * T_2 < t^* < T_2$ . Outside this range the performance slightly. For the sake of example, in Fig. 16, we show the probability for the fitness of the top ranked genome to belong to the top 50% values as a

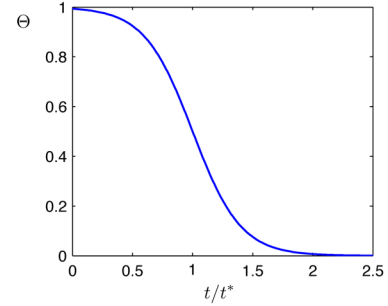


FIG. 14. The Fermi-Dirac function  $\Theta(t) = (1 + \exp(5 * (t/t^* - 1)))^{-1}$ .

function of the threshold parameter,  $t^*$ , for two different mutation rates.

We have also evaluated the performance of the algorithm for different sample sizes. In Fig. 17, we present the results for a set of parameters. We see that for samples smaller than  $n = 100$ , the performance decreases, whereas for higher samples sizes, the performance is similar to the results shown above for  $n = 200$ . For example, for a sample of size  $n = 30$ , and for parameters  $\mu = 5 * 10^{-3}$  and  $s = 2 * 10^{-3}$ , the probability for the fitness of the top ranked genome to belong to the top

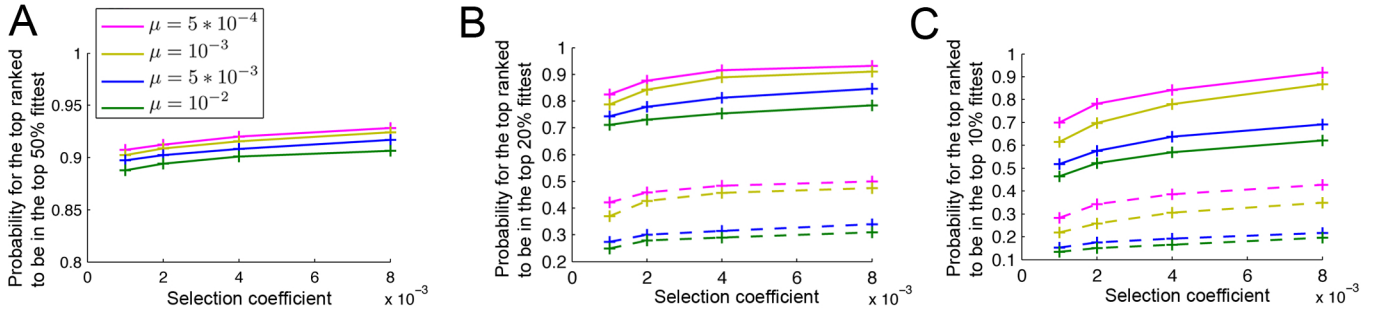


FIG. 15. Performance of the fitness ranking algorithm. (A) Probability for the fitness of the top ranked genome to belong to the top 50% fitness values of sampled genomes for a range of mutation rates and selection coefficients. (B) Probability for the fitness of the top ranked genome to belong to the top 20% fitness values shown using solid lines. The dashed lines show this probability for a randomly chosen genome (see the main text). (C) Probability for the fitness of the top ranked genome to belong to the top 10% fitness values. Sample size  $n = 200$ ,  $N = 64000$  and  $\epsilon = 0.1$  in all cases.

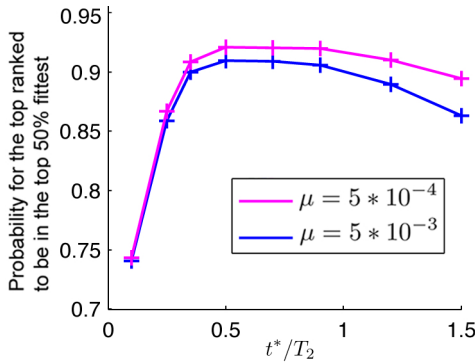


FIG. 16. Performance as a function of the threshold  $t^*$ .  $N = 64000$ ,  $\epsilon = 0.1$  and  $s = 2 * 10^{-3}$ .

50% values turns out to be around 0.84, compared to 0.9 for sample size of  $n = 200$ .

Another point is that, as sample size becomes smaller, the right tail of the fitness distribution (see Fig. 2A and

B) becomes under sampled. It has been shown that in similar models as the one we have considered here, the bulk of the fitness distribution can be approximated by a Gaussian profile [20]. For a Gaussian distribution, the probability of sampling a point with value of at least one (two)  $\sigma$  above the mean is around 0.15 (0.3). By inspecting the fitness profiles in Fig. 2A and B of the main text, as well as profiles shown in Fig. 6 of SI, we see that the frequency of clones with fitness more than one  $\sigma$  above the population average is around 0.1. This frequency for clones with fitness more than  $2\sigma$  above the population average is less than  $p = 0.05$ . In Fig. 17C, we see the ratio of the maximum fitness value in a sample of size  $n$  to the maximum fitness value that exist in the population. As expected, the larger the sample size, this ratio gets closer to one.

We have studied the performance of the algorithm in the presence of purifying selection. i.e. when  $\epsilon = 0$ . The results presented in Fig. 18 show that the algorithm performs well in this regime, similar to the regime of adaptation ( $\epsilon = 1$ ).

- 
- [1] Excoffier, L & Heckel, G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* **7**, 745–758.
  - [2] Hein, J, Schierup, M, & Wiuf, C. (2005) *Gene genealogies, variation and evolution: a primer in coalescent theory*. (Oxford University Press, USA).
  - [3] Kingman, J. (1982) The coalescent. *Stochastic processes and their applications* **13**, 235–248.
  - [4] Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* **123**, 585–595.
  - [5] Fu, Y & Li, W. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.
  - [6] Moya, A, Holmes, E, & González-Candelas, F. (2004) The population genetics and evolutionary epidemiology of rna viruses. *Nature Reviews Microbiology* **2**, 279–288.
  - [7] Neumann, G, Green, M, & Macken, C. (2010) Evolution of highly pathogenic avian h5n1 influenza viruses and the emergence of dominant variants. *Journal of General Virology* **91**, 1984–1995.
  - [8] Neher, R & Leitner, T. (2010) Recombination rate and selection strength in hiv intra-patient evolution. *PLoS computational biology* **6**, e1000660.
  - [9] Barrick, J, Yu, D, Yoon, S, Jeong, H, Oh, T, Schneider, D, Lenski, R, & Kim, J. (2009) Genome evolution and adaptation in a long-term experiment with escherichia coli. *Nature* **461**, 1243–1247.
  - [10] Sella, G, Petrov, D, Przeworski, M, & Andolfatto, P. (2009) Pervasive natural selection in the drosophila genome? *PLoS genetics* **5**, e1000495.
  - [11] Seger, J, Smith, W, Perry, J, Hunn, J, Kaliszewska, Z, La Sala, L, Pozzi, L, Rowntree, V, & Adler, F. (2010) Gene genealogies strongly distorted by weakly interfering mutations in constant environments. *Genetics* **184**, 529–



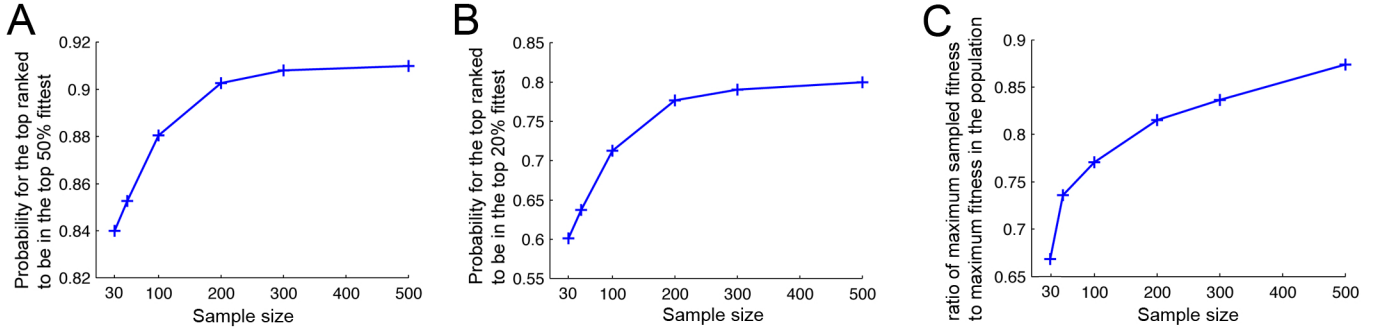


FIG. 17. Performance as a function of the sample size.  $N = 64000$ ,  $\epsilon = 0.1$ ,  $\mu = 5 \times 10^{-3}$  and  $s = 2 \times 10^{-3}$ .

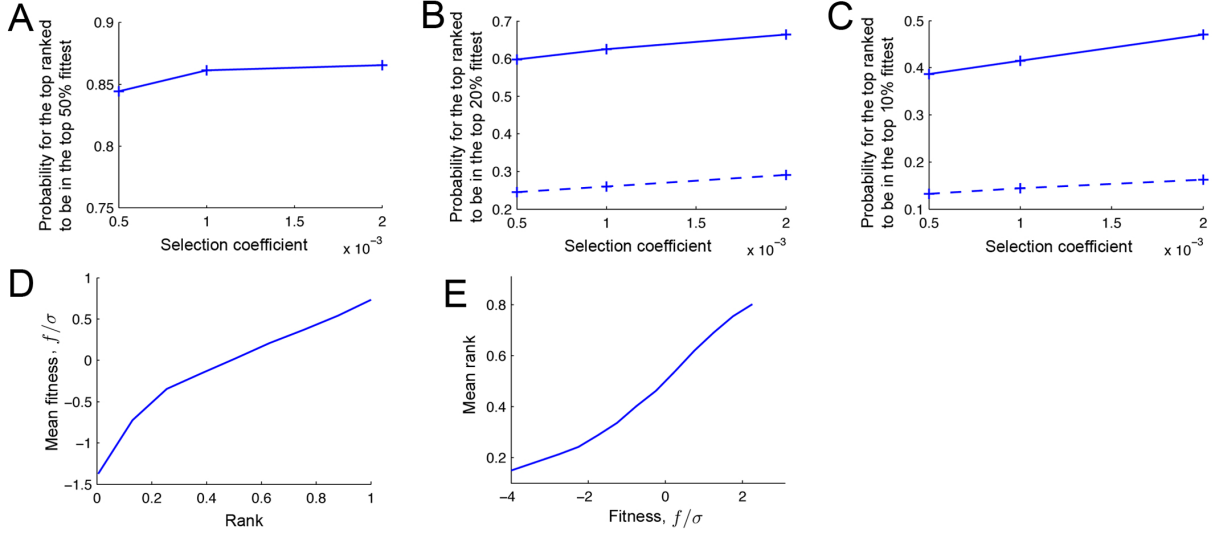


FIG. 18. Performance of the fitness ranking algorithm in the case of purifying selection. Sample size  $n = 200$ ,  $N = 32000$  and  $\mu = 5 \times 10^{-3}$  in all plots. (A) Probability for the fitness of the top ranked genomes to belong to the top 50% fitness values. (B) Probability for the fitness of the top ranked genomes to belong to the top 20% fitness values. The dashed line shows this probability for a randomly chosen genome. (C) Probability for the fitness of the top ranked genomes to belong to the top 10% fitness values. (D) Mean fitness as a function of the rank. Selection coefficient  $s = 10^{-3}$ . (E) Mean rank as a function of the fitness. Selection coefficient  $s = 10^{-3}$ .

- 545.
- [12] Merlo, L, Pepper, J, Reid, B, & Maley, C. (2006) Cancer as an evolutionary and ecological process. *Nature Reviews Cancer* **6**, 924–935.
- [13] Lenski, R, Rose, M, Simpson, S, & Tadler, S. (1991) Long-term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *American Naturalist* pp. 1315–1341.
- [14] Miralles, R, Gerrish, P, Moya, A, & Elena, S. (1999) Clonal interference and the evolution of rna viruses. *Science* **285**, 1745–1747.
- [15] Kao, K & Sherlock, G. (2008) Molecular characterization of clonal interference during adaptive evolution in asexual populations of saccharomyces cerevisiae. *Nature genetics* **40**, 1499–1504.
- [16] Lang, G, Botstein, D, & Desai, M. (2011) Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* **188**, 647–661.
- [17] Maia, L, Colato, A, & Fontanari, J. (2004) Effect of selection on the topology of genealogical trees. *Journal of theoretical biology* **226**, 315–320.
- [18] Tsimring, L, Levine, H, & Kessler, D. (1996) Rna virus evolution via a fitness-space model. *Physical review letters* **76**, 4440–4443.
- [19] Rouzine, I, Wakeley, J, & Coffin, J. (2003) The solitary wave of asexual evolution. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 587.
- [20] Desai, M & Fisher, D. (2007) Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759.
- [21] Rouzine, I, Brunet, É, & Wilke, C. (2008) The traveling-wave approach to asexual evolution: Muller’s ratchet and speed of adaptation. *Theoretical population biology* **73**, 24–46.
- [22] O’Fallon, B, Seger, J, & Adler, F. (2010) A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Molecular biology and evolution* **27**, 1162–1172.
- [23] Walczak, A, Nicolaisen, L, Plotkin, J, & Desai, M. (2012)

- The structure of genealogies in the presence of purifying selection: A fitness-class coalescent. *Genetics* **190**, 753–779.
- [24] Sniegowski, P & Gerrish, P. (2010) Beneficial mutations and the dynamics of adaptation in asexual populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 1255–1263.
  - [25] Goyal, S, Balick, D, Jerison, E, Neher, R, Shraiman, B, & Desai, M. (2011) Rare beneficial mutations can halt muller’s ratchet. *Arxiv preprint arXiv:1110.2939*.
  - [26] Bolthausen, E & Sznitman, A. (1998) On ruelle’s probability cascades and an abstract cavity method. *Communications in mathematical physics* **197**, 247–276.
  - [27] Brunet, E, Derrida, B, Mueller, A, & Munier, S. (2007) Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. *Physical Review E* **76**, 041104.
  - [28] Berestycki, N. (2009) Recent progress in coalescent theory. *Ensaio Matemáticos* **16**, 1–193.
  - [29] Desai, M, Walczak, A, & Fisher, D. (2012) Genetic diversity and the structure of genealogies in rapidly adapting populations. *Arxiv preprint arXiv:1208.3381*.
  - [30] Neher, R & Hallatschek, O. (2012) Genealogies of rapidly adapting populations. *Arxiv preprint arXiv:1208.3185*.
  - [31] Neher, R, Shraiman, B, & Fisher, D. (2010) Rate of adaptation in large sexual populations. *Genetics* **184**, 467–481.
  - [32] Smith, A, Heisler, L, Onge, R, Farias-Hesson, E, Wallace, I, Bodeau, J, Harris, A, Perry, K, Giaever, G, Pourmand, N, et al. (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic acids research* **38**, e142–e142.
  - [33] Navin, N, Kendall, J, Troge, J, Andrews, P, Rodgers, L, McIndoo, J, Cook, K, Stepansky, A, Levy, D, Esposito, D, et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94.
  - [34] Gordo, I & Charlesworth, B. (2000) The degeneration of asexual haploid populations and the speed of muller’s ratchet. *Genetics* **154**, 1379–1387.
  - [35] Hill, W. G & Robertson, A. (1966) The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.
  - [36] Neher, R & Shraiman, B. (2011) Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* **188**, 975–996.
  - [37] Squires, R, Noronha, J, Hunt, V, García-Sastre, A, Macken, C, Baumgarth, N, Suarez, D, Pickett, B, Zhang, Y, Larsen, C, et al. (2011) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses*.
  - [38] Plotkin, J, Dushoff, J, & Levin, S. (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proceedings of the National Academy of Sciences* **99**, 6263.
  - [39] Broman, K & Sen, S. (2009) A guide to qtl mapping with r/qtl (statistics for biology and health). *Recherche* **67**, 02.
  - [40] Durbin, R. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. (Cambridge Univ Pr).
  - [41] Felsenstein, J. (2004) Inferring phylogenies. *Sunderland, Massachusetts: Sinauer Associates*.
  - [42] Kingman, J. (1982) On the genealogy of large populations. *Journal of Applied Probability* pp. 27–43.
  - [43] Derrida, B & Peliti, L. (1991) Evolution in a flat fitness landscape. *Bulletin of mathematical biology* **53**, 355–382.
  - [44] Wakeley, J et al. (2008) Coalescent theory.